Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Neural Machine Translation enhanced with paraphrasing techniques

Degree's Thesis
Science and Technologies of Telecommunications Engineering

Author: Brian Tubay Álvarez
Advisors: Marta Ruiz Costa-Jussà

Universitat Politècnica de Catalunya (UPC)
2017 - 2018

Abstract

Deep Learning algorithms have a big impact in such areas as speech and image recognition or natural language processing. Machine Translation included. In recent years Neural Machine Translation (NMT) models have reached state-of-the art in this task of translating from a source language into another target language.
This project is developed with the aim of learning about the latest NMT architecture, The Transformer[1], which has become in the current state-of-the-art. This architecture out-stands for being the first NMT model that relies entirely on self-attention to compute representations of its inputs and outputs without Recurrent Neural Networks.
Despite of being developed recently, it has been used in diverse tasks. In this project, this model is implemented to develop a translator of biomedical texts. Biomedical field has the peculiarity that it does not count with huge databases of translated language pairs. In the aim of solving this problem, a multilingual translation system has been implemented: Romance languages like Spanish, French and Portuguese and their translation to English compose a large and unique dataset.
With the objective of evaluating these systems, the student has enrolled in the Biomedical Translation Task (WMT18). As the results are not published yet, for this bachelor's thesis, WMT17 datasets have been used to tests the translation systems.

Resum

Els algoritmes basats en Deep Learning ha suposat un gran impacte en àrees com el reconeixement de la parla i imatges o el processament del llenguatge natural. Traducció automàtica inclosa. Als darrers anys, els models de traducció automàtica basats en xarxes neuronals (NMT) han assolit l'estat del art en aquesta àrea de traduir un idioma font a un de destí.
Aquest projecte es desenvolupa amb la fi d'aprendre sobre la darrera arquitectura NMT, el Transformer[1], que s'ha convertit en l'actual estat de l'art. Aquesta arquitectura destaca per ser el primer model de NMT que es basa únicament en self-attention per obtenir representacions de les seves entrades i sortides sense xarxes recurrents.
Tot i haver estat desenvolupat fa poc, el seu ús és extens en diverses tasques. En aquest projecte en concret, el model s'utilitza per desenvolupar un traductor de texts biomèdics. El camp biomèdic té la peculiaritat que no compta amb grans datasets de traduccions d'idiomes. Amb l'objectiu de solucionar aquest problema, un sistema de traducció multilingüe és implementat: Llengües romàniques com el castellà, el francès i el portuguès, amb la seva traducció a l'anglès, composen un gran i únic data set.
Amb l'objectiu d'avaluar els traductors, s'ha participat en la Biomedical Translation Task (WMT18). Donat que els resultats encara no han estat publicats, per a aquesta memòria s'han fet servir els datasets del WMT17 per analitzar els sistemes de traducció.

Resumen

Los algoritmos basados en Deep Learning han supuesto un gran impacto en áreas como el reconocimiento del habla y imágenes o el procesamiento del lenguaje natural. Traducción automática incluida. En los últimos años, los modelos de traducción automática basados en redes neuronales (NMT) han alcanzado el estado del arte en esta tarea de traducir de un idioma fuente a un idioma destino.

Este proyecto se desarrolla con el fin de aprender acerca de la última arquitectura NMT, el Transformer[1], el cual se ha convertido en el actual estado del arte. Esta arquitectura destaca por ser el primer modelo de NMT que se basa únicamente en self-attention para obtener representaciones de sus entradas y salidas sin redes recurrentes.

A pesar de haber sido desarrollado recientemente, ya se ha utilizado en diversas tareas. En este proyecto en concreto, el modelo se utiliza para desarrollar un traductor de textos biomédicos. El campo biomédico tiene la peculiaridad que no cuenta con enormes datasets de traducciones de idiomas. Con el fin de solventar este problema, un sistema de traducción multilingüe es implementado: Lenguas románicas como el español, francés y portugués y su traducción al inglés componen un gran y único dataset.

Con el objetivo de evaluar los traductores, se ha participado en la Biomedical Translation Task (WMT18). Dado que los resultados no han sido publicados aún, para esta memoria se han utilizado los datasets de WMT17 para analizar los sistemas de traducción.

Acknowledgements

First of all I want to thank my supervisor Marta Ruiz, for the continuous help and support through all the project process. Thanks for giving me a little of your valuable time.

I specially want to thank my parents for all the efforts they done to get me here. And to my brother and sisters, who altogether are my life.

I would also to thanks Maria, Javier and Alberto, my supervisors at CaixaBank and Everis, for allow me to develop my autonomous learning in this field in work hours and help me.

And finally thank you, Emma, for always being by my side, even in almost every word of this document.

Revision history and approval record

Revision	Date	Purpose
0	10/06/2018	Document creation
1		Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Brian Tubay Álvarez	btubay@gmail.com
Marta Ruiz Costa-Jussa	marta.ruiz@upc.edu

Written by:		Reviewed and approved by:	
Date	10/06/2018	**Date**	-
Name	Brian Tubay	**Name**	Marta Ruiz
Position	Project Author	**Position**	Project Supervisor

Contents

1 Introduction **10**

1.1 Statement of purpose . . . 10

1.2 Requirements and specifications . . . 10

1.3 Methods and procedures . . . 10

1.4 Work Plan . . . 10

1.4.1 Work Packages . . . 11

1.4.2 Gantt Diagram . . . 11

1.5 Incidents and Modification . . . 12

2 State of the art **13**

2.1 Natural Language Processing and Machine Translation . . . 13

2.2 Deep Learning . . . 13

2.3 Neural Machine Translation . . . 14

3 Methodology **16**

3.1 The Transformer . . . 16

3.2 Romance Languages Training Corpus . . . 17

4 Implementation **18**

4.1 Databases . . . 18

4.2 Data Pre-Processing . . . 18

4.2.1 Translation of rare words . . . 19

4.3 Framework and Parameters . . . 19

5 Evaluation **20**

5.1 Single-language Translation . . . 20

5.2 Multiple-language Translation . . . 20

5.3 Results 20
5.3.1 Examples 21

6 Conclusions **23**

List of Figures

1.1 Gantt Diagram of the Degree Thesis 11

2.1 Structure of a perceptron 13

2.2 Structure of a MultiLayer Perceptron 14

2.3 Diagram of a Recurrent Neural Network 15

3.1 Simplified diagram of the Transformer model 17

List of Tables

4.1 Datasets provided by WMT18 18

5.1 Trained systems results for WMT17 20

5.2 Results for fr2en with external testset 21

5.3 Spanish to English examples for WMT18 21

5.4 Portuguese to English examples for WMT18 22

5.5 French to English examples for WMT18 22

Chapter 1

Introduction

1.1 Statement of purpose

The main goal of the project is to develop a Machine Translation system using Deep Learning Techniques. The idea is to develop the best system dealing with the problem of having short corpora.

The selected architecture is a recent proposal from the Google Research Team: The Transformer[1]. A Neural Machine Translation System that relies entirely on self-attention to compute representations of its input and output without using Recurrent Neural Networks, achieving much shorter training times and better results than existing state of the art in this area. To deal with short corpora problem, a system of Multi-Language romance languages is applied.

1.2 Requirements and specifications

As one of the main programming languages in machine learning nowadays, the code of this project has been developed in Python 3.6. Although the Transformer was originally developed in Tensorflow, the Deep Learning framework selected was PyTorch, given the motivation of the student to learn about it.

All the software has been launched in a cluster of 8 servers from the TSC department of the UPC, each with 2 Intel® Xeon® E5-2670 v3 2,3GHz 12N processors, and a total of 16 NVIDIA GTX Titan X GPUs. Each GPU has 12GB of memory and 3072 CUDA Cores.

1.3 Methods and procedures

The project's main idea was originally proposed by my supervisor, who had participated before in this type of competitions and encouraged me to do the same.

Transformer implementation it is not simple, for this reason it is based in the modules provided by the open-source toolkit Open-NMT[1], which is specialized in Neural Machine Translation tools.

1.4 Work Plan

The project followed the originally established work plan, with a few exceptions and modifications addressed in Section 1.5

[1] http://opennmt.net/OpenNMT-py/

1.4.1 Work Packages

- WP 1: Project propose and work plan
- WP 2: Information research
- WP 3: Project development
- WP 4: Critical Review
- WP 5: Biomedical Translation Task
- WP 6: Final Report
- WP 7: Oral Presentation

1.4.2 Gantt Diagram

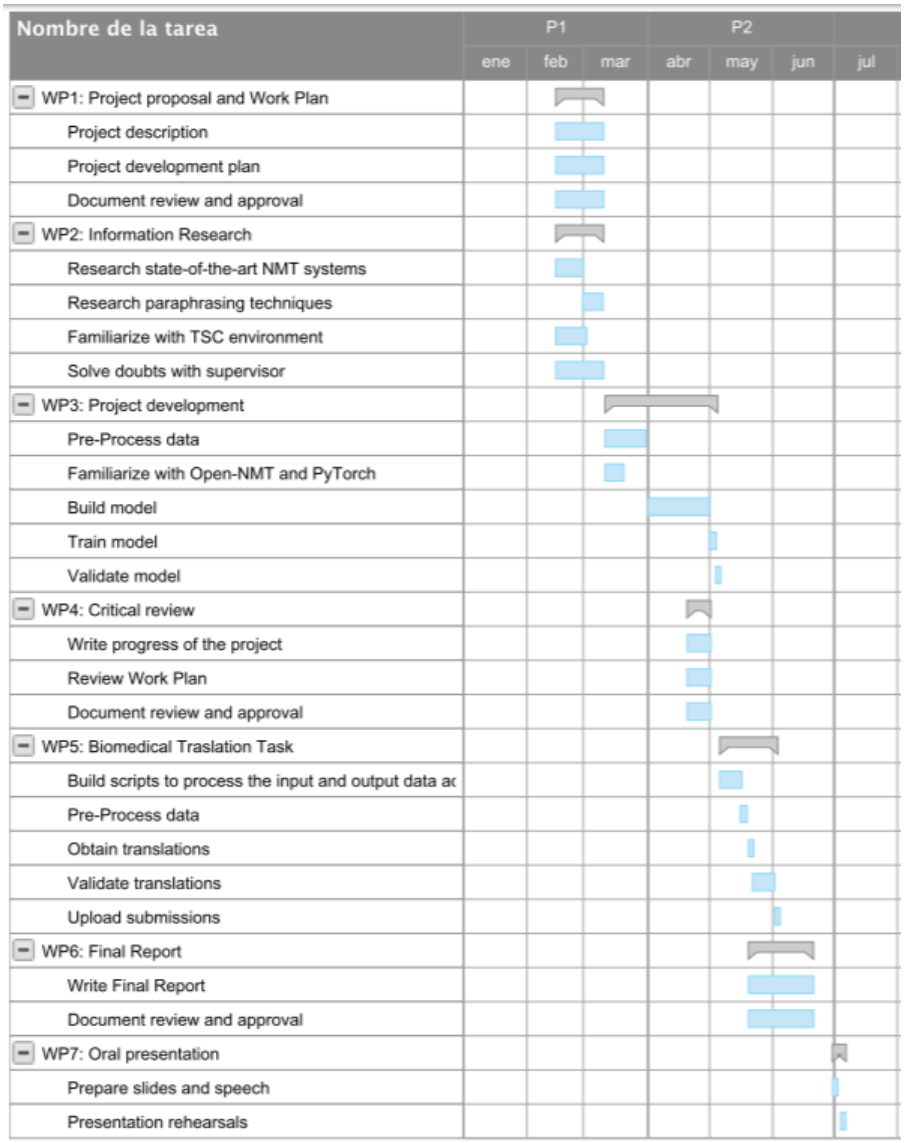

Figure 1.1: Gantt Diagram of the Degree Thesis

1.5 Incidents and Modification

The initial project was conceived to use paraphrasing techniques in order to solve short corpora problem. These techniques were researched and nearly applied to the project, but Biomedical Translation Task datasets do not contain the same phrases for the different language pairs, making impossible to apply pivoting methods [6].
As explained in Section 3.2, to solve this problem a system is trained with a common corpora composed by romance languages, creating an unique and common vocabulary.

Chapter 2

State of the art

Machine Translation has reached its state of the art based on Artificial Intelligence (AI) solutions. This chapter explains bases of Deep Learning and a wide vision of Neural Machine Translation models evolution through the years.

2.1 Natural Language Processing and Machine Translation

Natural Language Processing (NLP) is a research area of Artificial Intelligence (AI) that helps computers understand and manipulate human language. Machine Translation (MT) is a subfield NLP. It is the task of automatically convert one human language into another, preserving the meaning but also producing a correct structured text[1].

2.2 Deep Learning

Deep Learning is an area of Machine Learning that was inspired by the structure and function of biological neurons in the brain. The mathematical representation of the biological network in Deep Learning is the perceptron, being the basic unit of the system. These units are connected to each other and compose the neural network, being able to compute an output given the data input by decomposing it in different representations in order to identify diverse characteristics. Said simple, Neural Networks are designed to recognize patterns.

$$output = f(\sum_i x_i * w_i + b) \tag{2.1}$$

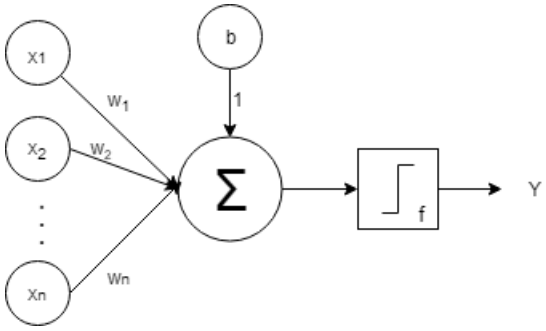

Figure 2.1: Structure of a perceptron

In fact, perceptron is an algorithm for learning a binary classifier. Vector input (x) is weighted (w) and biased (b) to finally being mapped to an output through an activation function (f).

[1] https://nlp.stanford.edu/projects/mt.shtml

The optimal values of weights and bias are computed using gradient descent techniques, in order to find the minimum distance between output of the perceptron and the actual labeled output desired.

Perceptron operates as a linear discriminant, then every unit can linearly separate inputs into two classes. So, it is possible to emulate basic linear operations as AND or OR, but no XOR, which represents a non-linear separable problem. These problems are solved with a neural network composed by multiple layers of perceptron, called MultiLayer Perceptron (MLP) or Feedforward Neural Network. Its structure is composed by an input layer, one or more hidden layers and an ouput layer. Input or visible layer provide the data to the network. It just simply pass the input into the next layer, so it is not composed by neurons with activation function as described before. Hidden layers learn multiple representations of the data. Output layer is responsible for outputting values. Its strong resides in its activation function, because it depends on the type of problem that is modeling.

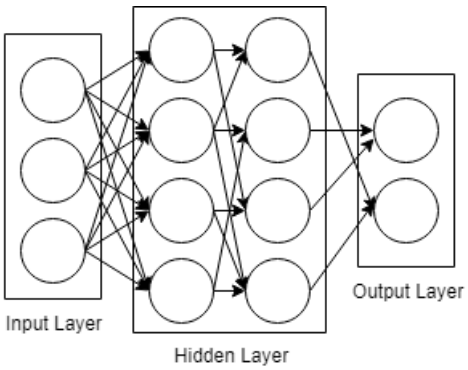

Figure 2.2: Structure of a MultiLayer Perceptron

2.3 Neural Machine Translation

Modern methods of Neural Machine Translation use an architecture based on the Encoder-Decoder model. The encoder takes the sequence of words in the source sentence (variable-lenght input), projects it into a fixed-length vector and passes this vector to the decoder. The decoder projects the vector into the original space of symbols, sequentially outputting the target sentence word for word.

Traditionally, both encoder and the decoder were composed of Recurrent Neural Networks (RNNs). RNN have the ability to retain information from previous data as a temporal memory. In fact, they can be viewed as a concatenation of the same unit in different time steps (Figure 2.3), where each one computes its output with an input and the output from the previous one. But RNN can only retain recent information from a sequence, performing well if the elements under the relational study are near to each other.

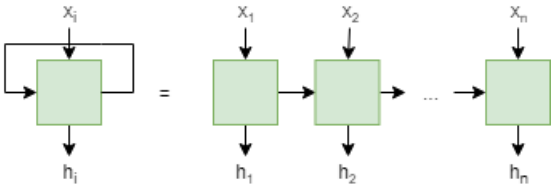

Figure 2.3: Diagram of a Recurrent Neural Network

Long Short Term Memory (LSTM) [4] units are a special type of RNN which perform well with long-term dependencies due to its internal structure is based in different types of gates. These gates are a way to optionally let information pass through. A variant of LSTMs are the Gate Recurrent Unit (GRU)[8], which have a simple structure and are computationally more efficient and faster to train.

Great models had been designed based in this architecture[5]. But it still had a problem. As the size of the sentences increases, a larger quantity of information have to be encoded in the fixed-length vector that the encoder passes the decoder, increasing the lost of information in this process. A solution to this problem is to allow the decoder to "attend" to the most relevant words of the source sentence in each step of the decoding process. This is called the Attention Mechanism. [3]

Chapter 3

Methodology

This chapter explains the encoder/decoder architecture used in this project: The Transformer[1].

3.1 The Transformer

The Transformer is the first Neural Machine Translation model relying entirely on self-attention to compute representations of its input and output without using RNNs or CNNs. It was proposed by Google team as a new state of the art in NMT. And not only for this purpose, they proved that it can be accomplish other task as English constituency parsing [9].

As explained before, RNNs read one word at a time, forcing itself to perform multiple steps to make decisions that depends on words that are far away from each other. But it has been demonstrated that the more steps required, the harder it is to the network to learn how to make these decisions. In addition, if the sequential nature of the RNNs are taking into account, results that it difficult to fully take advantage of modern computing devices such as Tensor Processing Units (TPUs) or Graphics Processing Units (GPUs) which outperform in parallel processing. The Transformer is an encoder-decoder model conceived to solve this problems.

The encoder is composed of three stages. In the first stage input words are projected into a vectorial space (embedded vectors). Unlike in RNNs, there is no information of the position of the words in the sentence, a positional encoding is added to the embedded input vectors[1]. The second stage is a multi-head self-attention. Instead of computing single attention, this stage compute multiple attention over the source and realize a weighted sum between them[2]. Finally a position-wise fully connected feed-forward network is used, which consists of two linear transformations with a ReLU activation[11] in between.

The decoder operates similarly, but generates one word at a time, from left to right. It is composed of five stages. The first two are similar to the encoder: embedding and positional encoding and a masked multi-head self-attention, which unlike in the encoder, forces to attend only to past word. The third stage is a multi-head attention that not only attends to these past words, but also to the final representations generated by the encoder. The fourth stage is another position-wise feed-forward network. Finally, the softmax layer allows to map target word scores into target word probabilities. For more specific details about the architecture see the original paper [1]

[1]Without positional encodings, the output of the multi-head attention network would be the same for the sentences "I love you more than her" and "I love her more than you".

[2]Multi-head attention is composed by different submodules of scaled dot-product attentions with different linear projections.

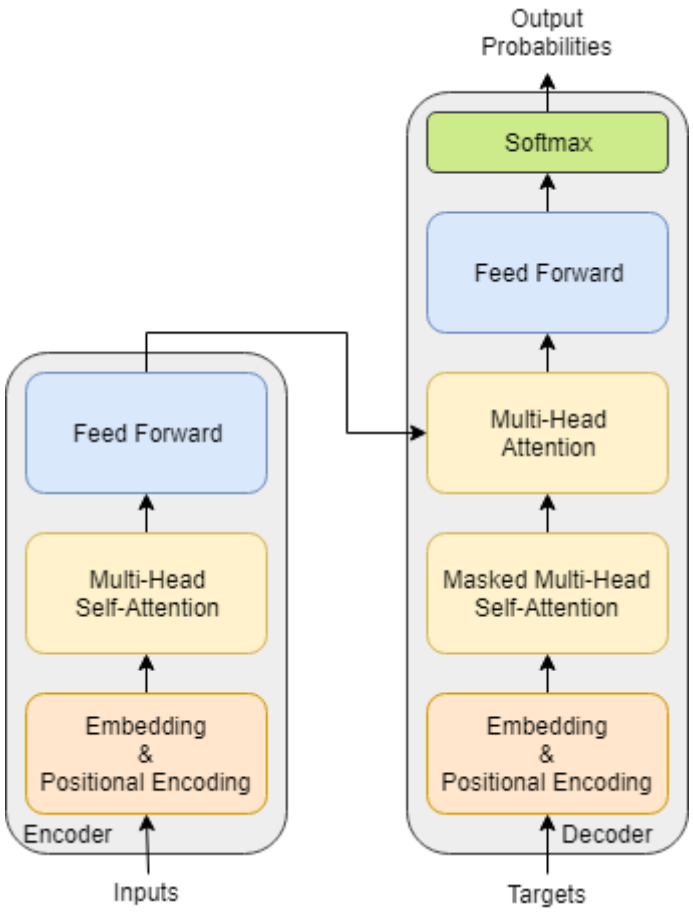

Figure 3.1: Simplified diagram of the Transformer model

3.2 Romance Languages Training Corpus

Training a Deep Learning model with short corpora could represent a big problem to guarantee desired results. This problem applies in our case, due to the nature of the Biomedical Translation Task. The details of the number of sentences provided by the Task are described in Section 4.1

With this inconvenient, the supervisor proposed the creation of a unique corpus composed by sub-corpus of romance languages such as Spanish, Portuguese and French as the input language and their English translation as the target language. The idea was to prove that the model trained with this corpus will outperform each model trained with each sub-corpus individually. This hypothesis has its origin in [12], where the authors build a multi-source MT model to maximize the target probability (English) given French and German sources.

Chapter 4

Implementation

In this chapter the databases used and its fine tuning (preprocessing) to be trainable with better results. Next, the implementation of the model will be exposed, explaining which parameters were selected and why and how models were trained. Finally, it will be explained how the results were tested and which was the criterion to interpret the results.

4.1 Databases

As the project is presented to the Biomedical Translation Task (WMT18)[1] the databases used to train the model are the ones provided by the task for the languages translation pairs selected: es2en, fr2en and pt2en. These datasets are mainly from Scielo and Medline databases.

Training	Scielo	Medline	Total
es2en	713.127	285.358	998.485
fr2en	9.127	612.645	621.772
pt2en	634.438	74.267	708.705
all2en	1.356.692	972.270	2.328.962

Table 4.1: Datasets provided by WMT18

Validation datasets for three languages were obtained from the *Khresmoi development data*[2], as recommended at the task. Each validation dataset contains 500 pairs of phrases.

The selected test datasets were the ones provides by the task for the last year competition (WMT17)[3]. The reason is obviously to compare the quality of the Machine Translation system with the best system in that edition. This comparison is detailed in Chapter 5

4.2 Data Pre-Processing

In order to train the model, a pre-processing is required. Pre-processing relied in three basic steps: Tokenization, Truecasing and Cleaning the corpus. The scripts to realize these steps are provided by Moses[4], a statistical machine translation system.

- Tokenization: Spaces are inserted between words and punctuation. This is done because data is feed in the net by tokens, which are defined by being separated by spaces. So, punctuation will be considered a token such as a word.

[1] http://www.statmt.org/wmt18/biomedical-translation-task.html
[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122
[3] http://www.statmt.org/wmt17/biomedical-translation-task.html
[4] http://www.statmt.org/moses/

- Truecasing: The initial words in each sentence are converted to their most probable casing (This probability are defined by a previous training). This helps reduce data sparsity.
- Cleaning: Limit sentences into a range of minimum and maximum number of tokens, comparing the two datasets (language to be translated and language translated).

4.2.1 Translation of rare words

Another step is realized before training. What happens if a word is out of the vocabulary (OOV) of the training corpus? It can not be translated, so the special unknown token (UNK) is obtained as translation. In the case of the available corpus for this task, which is limited, it is easy that this situation happens.

In order to minimize this effect, a system that works on subword units is used [10]. Words are segmented using Byte Pair Encoding (BPE) algorithm into subword units of different length, and a new vocabulary is built using frequent such units.

4.3 Framework and Parameters

As said before, Transformer architecture was developed by Google, who used its own software Tensorflow in order to develop it inside the well-known package tensor2tensor.

The Transformer model is very sensitive to hyperparameters. Hence, OpenNMT developers provide a set of parameters that mimic the Google original setup [1], replicating their WMT results, but in PyTorch. These parameters could be found in their website[5] [2].

[5] http://opennmt.net/OpenNMT-py/FAQ.html

Chapter 5

Evaluation

5.1 Single-language Translation

As detailed in 4.1, three language pairs were selected: Spanish to English (es2en), French to English (fr2en) and Portuguese to English (pt2en). Three systems were trained, one for each language pair. 14 epochs were required for the es2en system, with a training duration of 7 hours. 16 epochs were required for the fr2en, with a training duration of 9 hours. And for pt2en 17 epochs were required, with a training duration of 7 hours.

5.2 Multiple-language Translation

As explained before, in order to improve the results in each language pair, another system was trained. This system, because of be composed by three datasets together, needed 11 epochs and a training duration of 23 hours.

5.3 Results

In order to compare the quality of a translation, the competition uses the Bilingual Evaluation Understudy (BLEU)[7], the most extended method for this task. This method compares the similarity of machine translated sentences with the actual translation (a human translation), returning a similarity score. Open-NMT incorporates a script (provided by the Moses Toolkit) that easily let to compute the BLEU score of a translation.

Best BLEU scores of the WMT17 for each language pairs are considered baseline. BLEU scores for each Machine Translation systems are the followings[1]:

System	es2en	pt2en	fr2en
Baseline	37.49	43.88	23.41
Single-Language	39.35	44.31	-
Multi-Language	40.11	45.55	-

Table 5.1: Trained systems results for WMT17

To compare the two systems in fr2en, a testset pair of Khresmoi development data were selected (1000 phrases). BLEU scores for this dataset are the followings:

[1]Testset for fr2en was not available in WMT17

System	fr2en
Single-Language	31.75
Multi-Language	38.31

Table 5.2: Results for fr2en with external testset

As we can see, Transformer architecture jointly with the neat pre-processing system outperform best results of WMT17. Results become even better with the system trained with the common corpus of romance languages, validating our purpose.

5.3.1 Examples

As explained, BLEU is a widely extended method for comparison of quality of translations. But in this kind of competition it is not the only one; a human validation also is performed. The results of this validation is not available yet, so in the following tables translation examples with both systems could be found.

Spanish to English

Original	Utilizando la base de datos Epistemonikos, la cual es mantenida mediante búsquedas realizadas en 30 bases de datos, identificamos seis revisiones sistemáticas que en conjunto incluyen 36 estudios aleatorizados pertinentes a la pregunta.
Spanish-system	Using the Epistemonikos database, which is maintained through searches in 30 databases, we identified six systematic reviews including 36 randomized studies relevant to the question.
Romance-system	Using the Epistemonikos database, which is maintained through searches in 30 databases, we identified six systematic reviews that altogether include 36 randomized studies relevant to the question.

Table 5.3: Spanish to English examples for WMT18

Portuguese to English

Original	Os resultados dos modelos de regressão mostraram associação entre os fatores de correção estimados e os indicadores de adequação propostos
Portuguese-system	Regression models showed an association between estimated correction factors and the proposed adequacy indicators.
Romance-system	The results of the regression models showed an association between the estimated correction factors and the proposed adequacy indicators.

Table 5.4: Portuguese to English examples for WMT18

French to English

Original	(Traduit par Docteur Serge Messier).
French-system	[Doctor Serge Messier].
Romance-system	[(Translated by Doctor Serge Messier)].

Table 5.5: French to English examples for WMT18

Chapter 6

Conclusions

The main objective of this project was to develop a translation system with latest state-of-the art techniques. The translation system is experimented in the biomedical domain, implying that low resources are available. For this reason, all the efforts were focused on this low-resources challenge.
It is pity that training datasets provided by Biomedical Translation Task were not the same for all languages, because it has prevented us to apply paraphrasing techniques to face the low-resource challenge [6]. Even so, another solution has been implemented with good results. It it demonstrated, then, that a common corpora build from related languages like romance languages outperforms trained models with single corpora, becoming a great solution if available. In fact, as mentioned before in 3.2, [12] already proved this hypothesis, but this project has tested not only with a different architecture but also it has been developed for another domain where source languages and target language do not belong to the same language family.

This project had a personal objective for me which was learning all I could of this field that is Deep Learning focused in Natural Language Processing. And I consider that this goal of the thesis has been accomplished for sure. I remember when I came to my supervisor fascinated with the famous paper of Sutskever et al.[5]. She told me that that paper was good, but also outdated, and gave me a lot of more updated papers to read. That was great.
After weeks of reading and solving doubts, and a few more weeks of battles with the data pre-processing, the main objective was finished and then improved with the Multilingual system. Now it only remains to know how well my systems performed in the WMT18.

Bibliography

[1] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. Attention is all you need. 2017.

[2] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. 2014.

[3] Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate. 2014.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, pages 1735–1780, 1997.

[5] Quoc V. Le Ilya Sutskever, Oriol Vinyals. Sequence to sequence learning with neural networks. 2014.

[6] Rico Sennrich Jonathan Mallinson and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, 2017.

[7] Todd Ward Kishore Papineni, Salim Roukos and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[8] Caglar Gulcehre Dzmitry Bahdanau Fethi Bougares Holger Schwenka Kyunghyun Cho, Bart van Merrienboer and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.

[9] Noam Shazeer Ashish Vaswani Niki Parmar Llion Jones Jakob Uszkoreit Lukasz Kaiser, Aidan N. Gomez. One model to learn them all. 2017.

[10] Alexandra Birch Rico Sennrich, Barry Haddow. Neural machine translation of rare words with subword units. 2015.

[11] Geoffrey E. Hinton Vinod Nair. Rectified linear units improve restricted boltzmann machines. In *27th International Conference on Machine Learning*, 2010.

[12] Barret Zoph and Kevin Knight. Multi-source neural translation. In *NAACL-HLT 2016*, pages 30–34, 2016.